Support Vector Methods for Functional Genomic Analysis

William Noble Grundy
Department of Computer Science
Columbia University

Outline

Gene functional classification using support vector machines.

- Learning from gene expression data.
- Learning from promoter region sequences.
- Learning from two types of data.

Acknowledgments

www.cs.columbia.edu/compbio

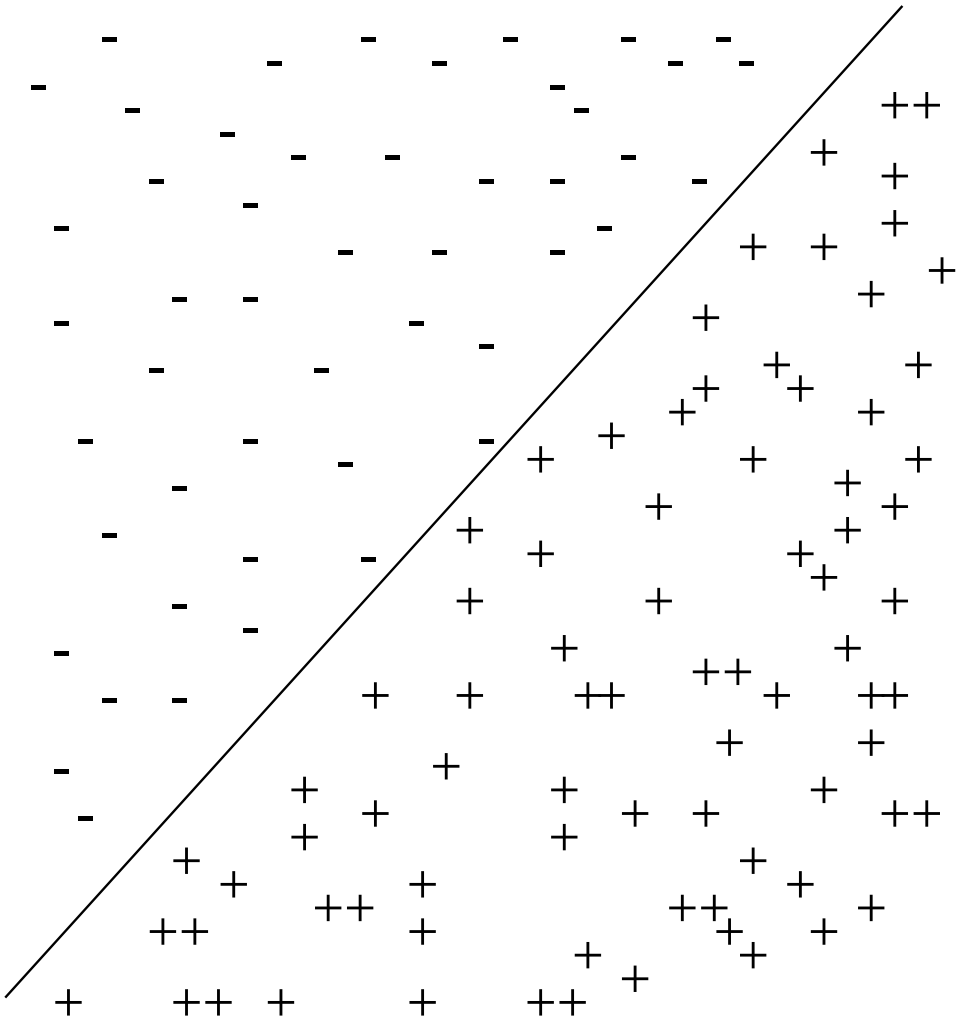
Promoter region analysis

- Paul Pavlidis, Columbia Genome Center
- Terry Furey, CS, UCSC
- Muriel Liberto, Biology, Columbia
- Prof. David Haussler, CS, UCSC

Heterogeneous data

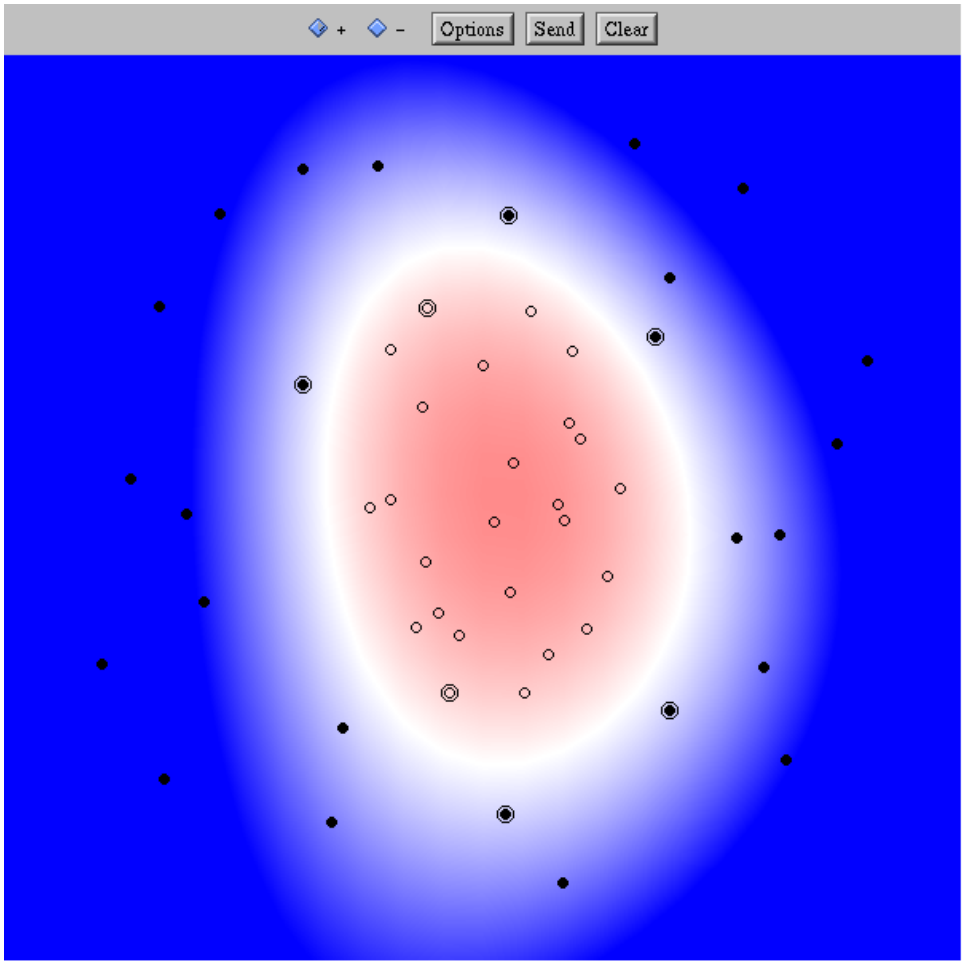
- Paul Pavlidis, Columbia Genome Center
- Jinsong Cai, Medical Informatics, Columbia
- Jason Weston, Barnhill Technologies

Separating hyperplane



- Each vector in the gene expression matrix may be thought of as a point in a 79-dimensional *input space*.
- A simple way to build a binary classifier is to construct a hyperplane separating class members from non-members in this space.
- This is the approach taken by perceptrons, also known as single-layer neural networks.

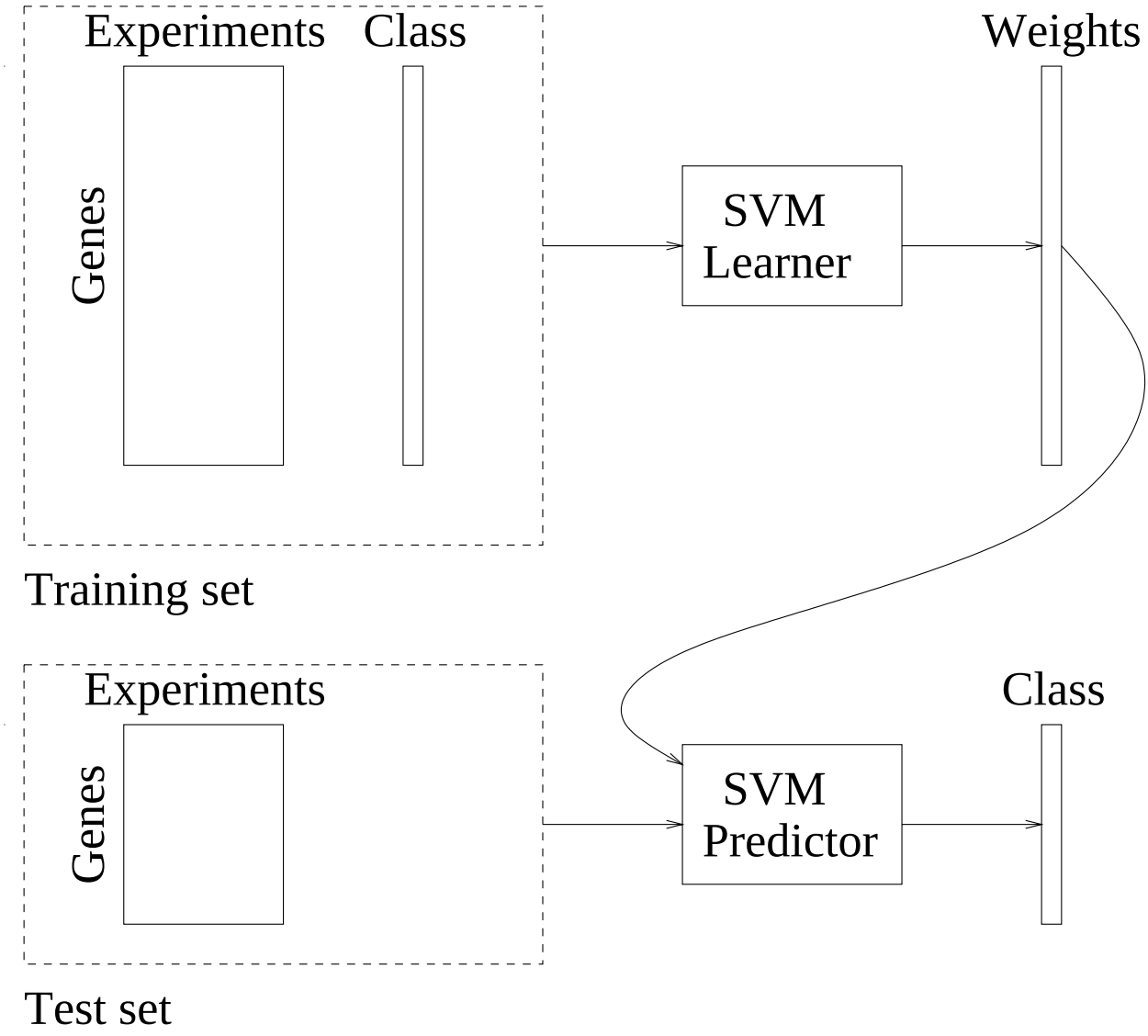
Gaussian decision boundary



$$K(X,Y) = exp\left(\frac{-||X-\bar{Y}||^2}{2\sigma^2}\right)$$

A radial basis kernel function yields a Gaussian decision boundary in the input space.

SVMs for gene functional classification

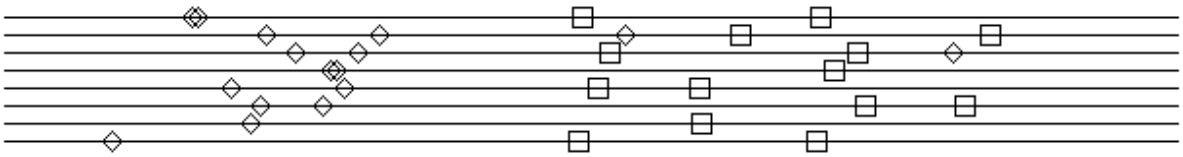


Classification performance

Method	TCA	Resp	Ribo	Prot	Hist	HTH
D-p 1 SVM	6	31	224	35	18	-56
D-p 2 SVM	9	39	229	48	18	-3
D-p 3 SVM	12	38	229	51	18	-1
Radial SVM	11	33	226	52	18	0
Parzen	6	18	220	39	14	-14
FLD	5	30	217	39	16	-14
C4.5	-7	8	169	33	16	-2
MOC1	-1	-4	164	26	10	-6

- Values reported are cost savings relative to the null procedure that classifies all examples as negatives.
- Cost is defined as the number of false positives plus twice the number of false negatives.

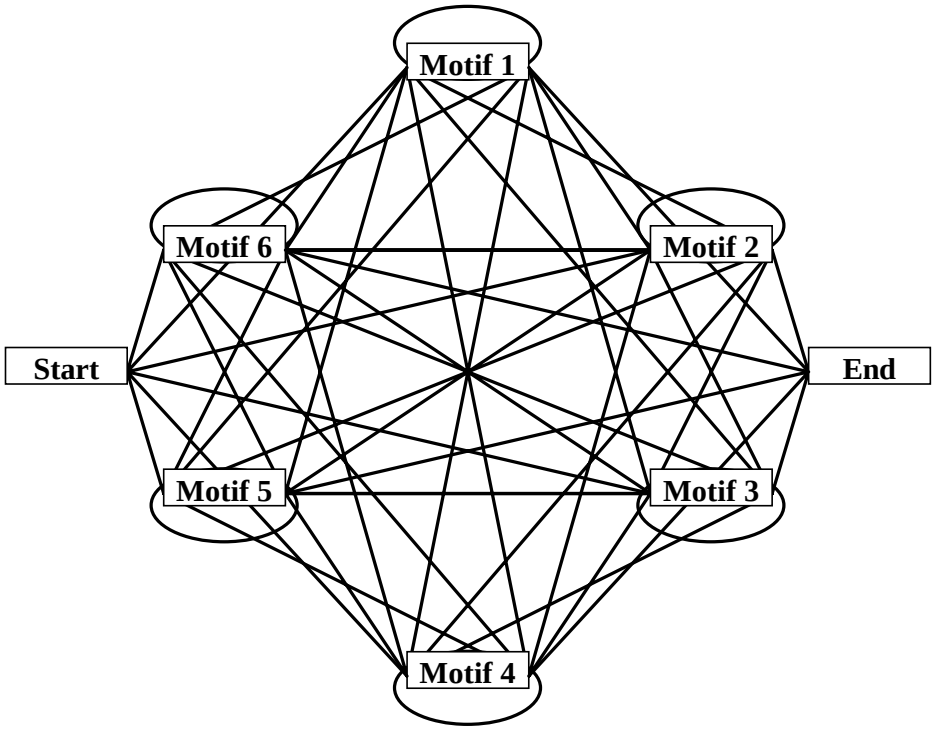
Promoter region analysis



Motif occurences in the nucleosomal promoters.

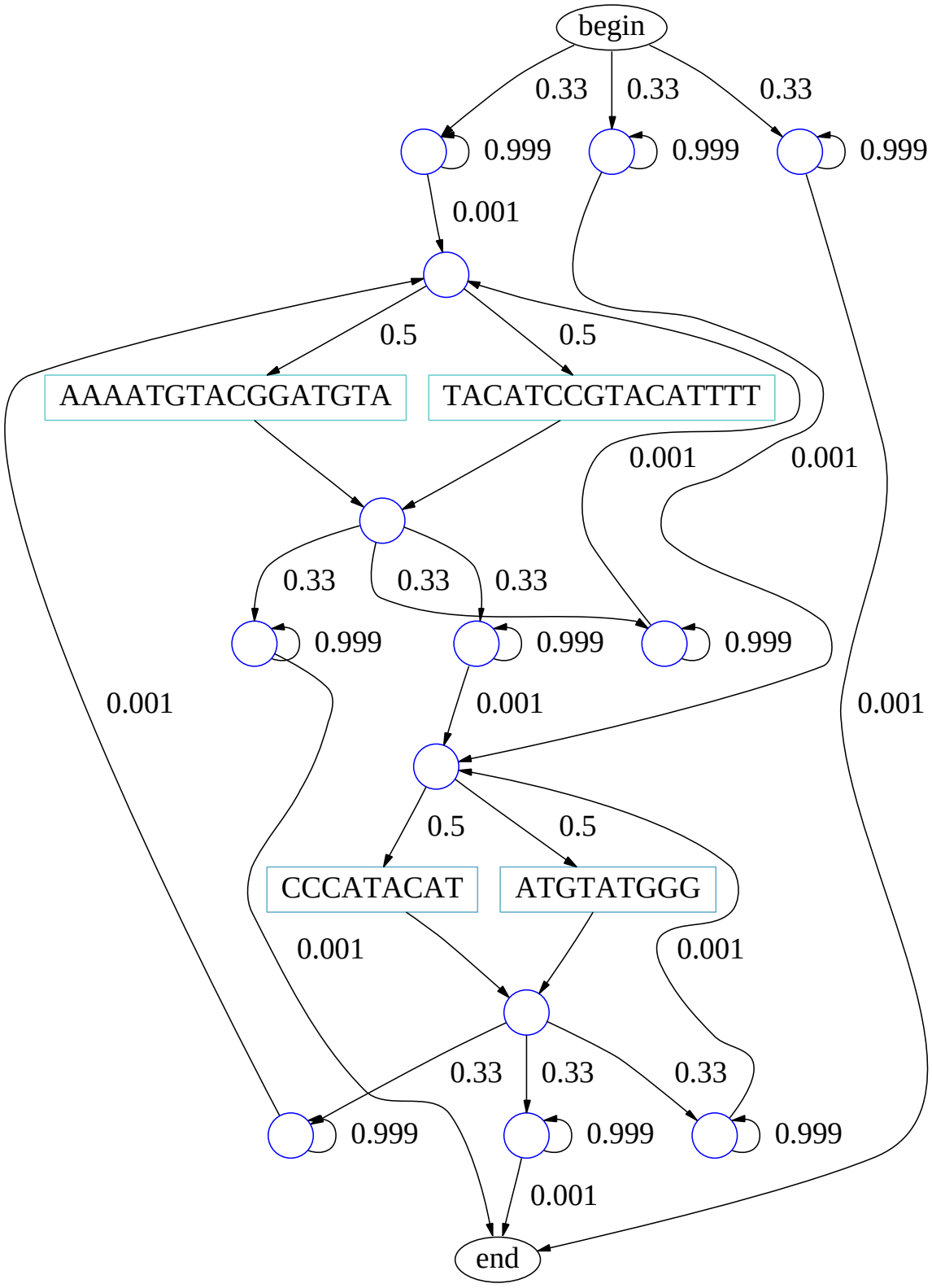
- Each line corresponds to a 1000-base pair nucleosomal promoter region.
- Boxes and diamonds represent motif occurrences.

Meta-MEME



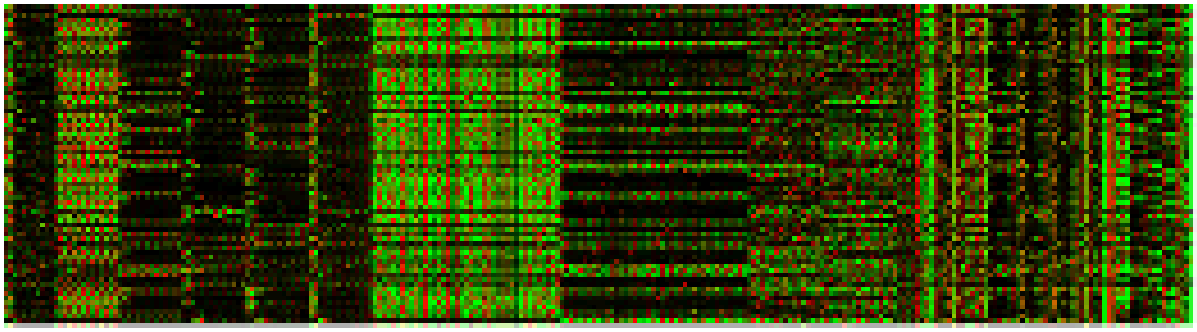
- Meta-MEME combines gapless motif models in a hidden Markov model framework.
- Meta-MEME models have fewer parameters than standard profile HMMs.
- The completed connected model topology allows for the repetition or shuffling of motifs or domains.

A model of ribosomal protein promoters

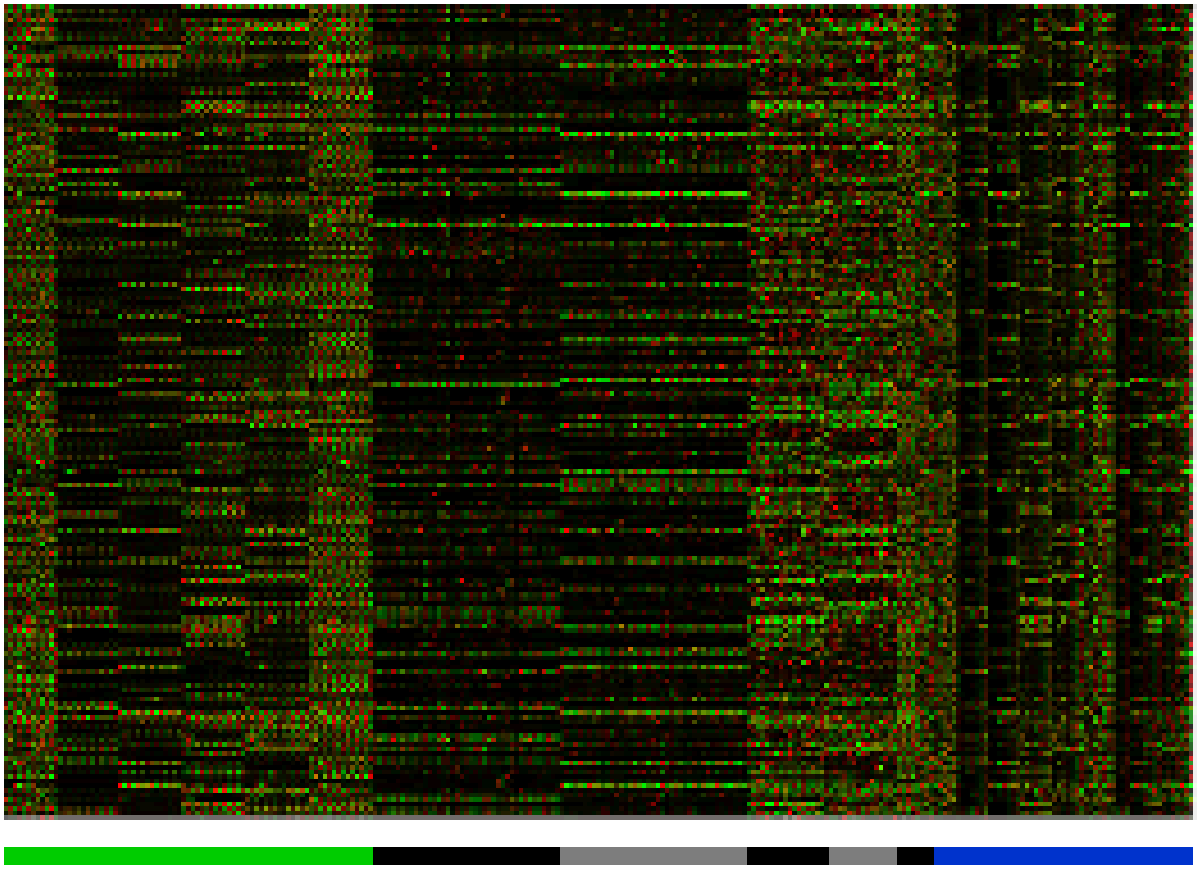


Visualizing Fisher score vectors

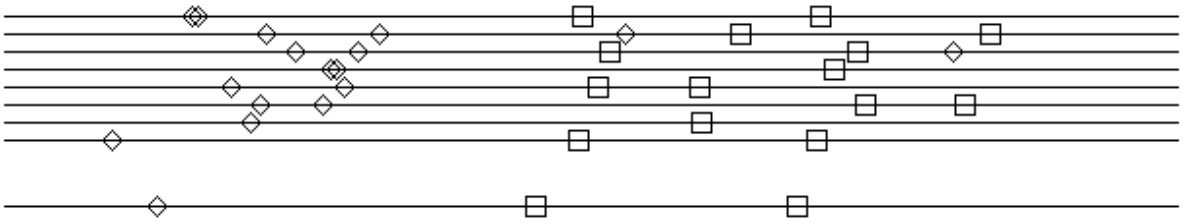
A



B



Nucleosomal prediction



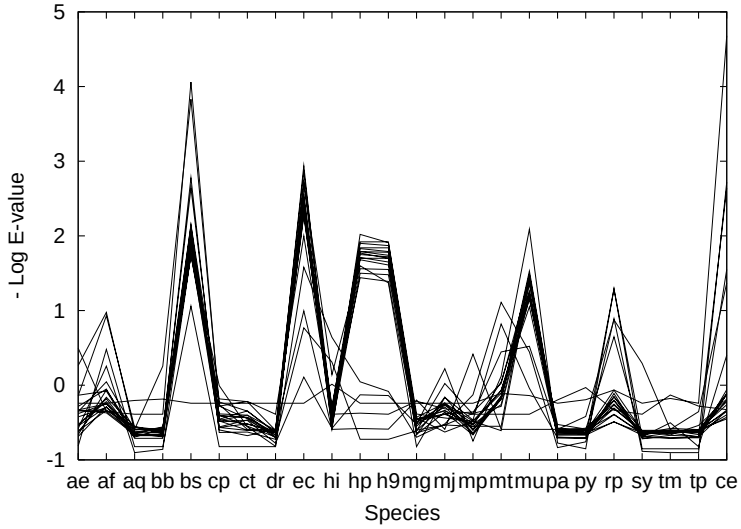
The final line is a promoter from a gene (YOR084W) identified by the Meta-MEME + SVM method.

Phylogenetic profiles

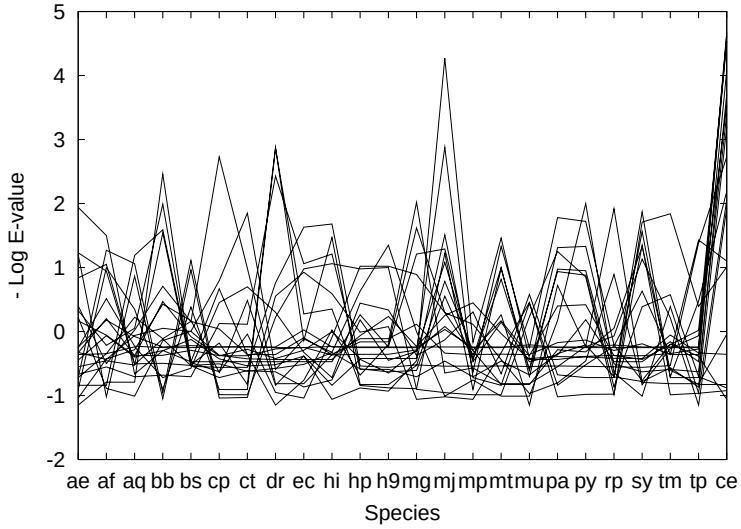
	Synechocystis sp.	M. genitalium	S. cerevisiae	H. pylori	M. pneumoniae	H. influenzae	M. jannaschi	E. coli
YAL001C	12.23	43.44	1.454	0.000	22.08	63.08	0.000	4.345
YAL002W	0.000	0.000	0.000	2.243	0.000	2.909	0.000	0.000
YAL003W	0.000	37.94	0.000	67.98	12.34	14.76	12.34	2.345
YAL005C	14.43	23.45	1.211	0.000	19.87	67.00	0.000	13.45

- For a given pair of genes, a similar pattern of inheritance across species may imply a functional link.
- Each profile entry is the negative log E-value of the top-scoring sequence from a BLAST search of a complete genome.
- Negative values (corresponding to E-values greater than 1) are set to zero.

Similar patterns of inheritance

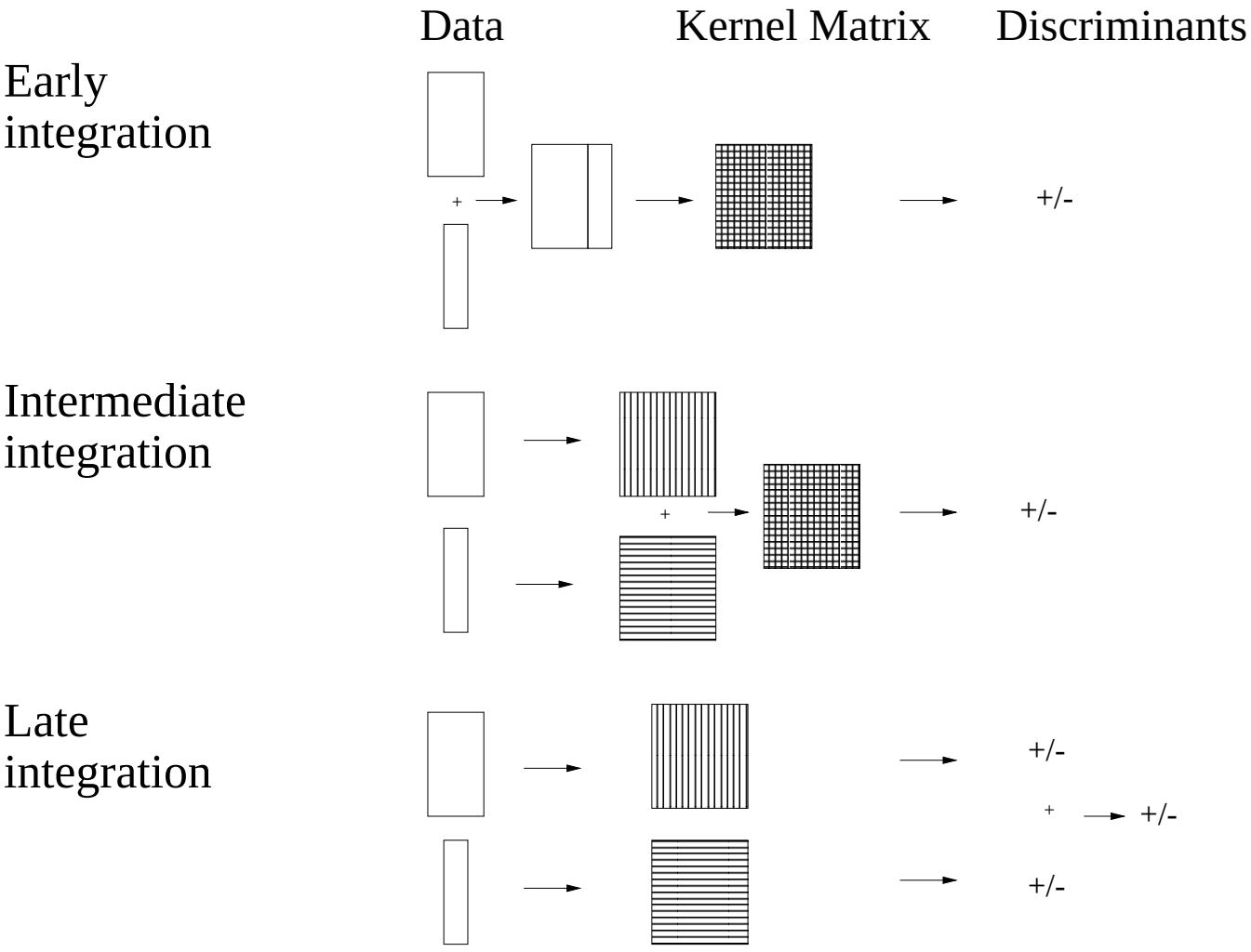


Phylogenetic profiles of 22 amino acid transporter genes.



Phylogenetic profiles of 22 randomly selected genes.

Learning from heterogeneous data



Comparison of data integration methods

Method	Cost savings	Best	Non-learnable
Gene expression	0.19 ± 0.02	10	4
Phylogenetic profiles	0.21 ± 0.04	12	6
Early integration	0.27 ± 0.03	17	3
Intermediate integration	0.31 ± 0.03	21	2
Late integration	0.24 ± 0.03	8	3

- 27 classes are included.
- Cost savings of 1 is perfect; 0 is comparable to classifying everything as negative.
- "Best": cost savings is within one standard deviation of the best cost savings.
- "Non-learnable": cost savings is within one standard deviation of zero.